Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences

Jolma, Arttu

2020-07

Jolma , A , Zhang , J , Mondragon , E , Morgunova , E , Kivioja , T , Laverty , K U , Yin , Y , Zhu , F , Bourenkov , G , Morris , Q , Hughes , T R , Maher III , L J & Taipale , J 2020 , ' Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences ' , Genome Research , vol. 30 , no. 7 , pp. 962-973 . https://doi.org/10.1101/gr.258848.119

http://hdl.handle.net/10138/319234
https://doi.org/10.1101/gr.258848.119

Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences

Arttu Jolma,[1,11] Jilin Zhang,[1,11] Estefania Mondragón,[2,11] Ekaterina Morgunova,[1] Teemu Kivioja,[3] Kaitlin U. Laverty,[4] Yimeng Yin,[1] Fangjie Zhu,[1] Gleb Bourenkov,[5] Quaid Morris,[4,6,7,8,9] Timothy R. Hughes,[4,6] Louis James Maher III,[2] and Jussi Taipale[1,3,10]

[1]*Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-171 77, Solna, Sweden;* [2]*Department of Biochemistry and Molecular Biology, Mayo Clinic Graduate School of Biomedical Sciences, Mayo Clinic College of Medicine and Science, Rochester, Minnesota 55905, USA;* [3]*Genome-Scale Biology Program, University of Helsinki, FI-00014, Helsinki, Finland;* [4]*Department of Molecular Genetics, University of Toronto, M5S 1A8, Toronto, Canada;* [5]*European Molecular Biology Laboratory (EMBL), Hamburg Unit c/o DESY, D-22603 Hamburg, Germany;* [6]*Donnelly Centre, University of Toronto, M5S 3E1, Toronto, Canada;* [7]*Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, M5S 3G4, Toronto, Canada;* [8]*Department of Computer Science, University of Toronto, M5S 2E4, Toronto, Canada;* [9]*Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA;* [10]*Department of Biochemistry, University of Cambridge, CB2 1QW, Cambridge, United Kingdom*

RNA-binding proteins (RBPs) regulate RNA metabolism at multiple levels by affecting splicing of nascent transcripts, RNA folding, base modification, transport, localization, translation, and stability. Despite their central role in RNA function, the RNA-binding specificities of most RBPs remain unknown or incompletely defined. To address this, we have assembled a genome-scale collection of RBPs and their RNA-binding domains (RBDs) and assessed their specificities using high-throughput RNA-SELEX (HTR-SELEX). Approximately 70% of RBPs for which we obtained a motif bound to short linear sequences, whereas ~30% preferred structured motifs folding into stem–loops. We also found that many RBPs can bind to multiple distinctly different motifs. Analysis of the matches of the motifs in human genomic sequences suggested novel roles for many RBPs. We found that three cytoplasmic proteins—ZC3H12A, ZC3H12B, and ZC3H12C—bound to motifs resembling the splice donor sequence, suggesting that these proteins are involved in degradation of cytoplasmic viral and/or unspliced transcripts. Structural analysis revealed that the RNA motif was not bound by the conventional C3H1 RNA-binding domain of ZC3H12B. Instead, the RNA motif was bound by the ZC3H12B's PilT N terminus (PIN) RNase domain, revealing a potential mechanism by which unconventional RBDs containing active sites or molecule-binding pockets could interact with short, structured RNA molecules. Our collection containing 145 high-resolution binding specificity models for 86 RBPs is the largest systematic resource for the analysis of human RBPs and will greatly facilitate future analysis of the various biological roles of this important class of proteins.

[Supplemental material is available for this article.]

The abundance of RNA and protein molecules in a cell depends both on their rates of production and degradation. The transcription rate of RNA and the rate of degradation of proteins are determined by DNA and protein sequences, respectively (Liu et al. 2016). However, most regulatory steps that control gene expression are influenced by the sequence of the RNA itself. These processes include RNA splicing, localization, stability, and translation, all of which can be regulated by RNA-binding proteins (RBPs) that specifically recognize short RNA sequence elements (Glisovic et al. 2008).

RBPs can recognize their target sites using two mechanisms: They can form direct contacts to the RNA bases of an unfolded RNA chain and/or recognize folded RNA structures (for reviews, see Draper 1999; Jones et al. 2001; Mackereth and Sattler 2012). These two recognition modes are not mutually exclusive, and the same RBP can combine both mechanisms in recognition of its target sequence. The RBPs that bind to unfolded target sequences are commonly assumed to bind to each base independently of the other bases, and their specificity is modelled by a simple position weight matrix (PWM) (Stormo 1988; Cook et al. 2011). However, recognition of a folded RNA sequence leads to strong positional interdependencies between different bases owing to base-pairing. In addition to the canonical Watson–Crick base pairs G:C and A:U, double-stranded RNA commonly contains also G:U base pairs and can also accommodate other noncanonical base-pairing configurations in specific structural contexts (Varani and McClain 2000).

It has been estimated that the human genome encodes approximately 1500 proteins that can associate with RNA (Gerstberger et al. 2014). Only some of the RBPs are thought to be sequence specific. Many RBPs bind only a single RNA species (e.g., ribosomal proteins) or serve a structural role in

[11]These authors contributed equally to this work.
Corresponding author: ajt208@cam.ac.uk
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.258848.119. Freely available online through the *Genome Research* Open Access option.

30:962–973 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/20; www.genome.org

ribonucleoprotein complexes or the spliceosome. As RNA can fold to complex three-dimensional structures, defining what constitutes an RBP is not simple. In this work, we have focused on identifying motifs for RBDs that bind to short sequence elements, analogously to sequence-specific DNA binding transcription factors. The number of such RBPs can be estimated based on the number of proteins containing one or more canonical RNA-binding protein domains (RBDs). The total number is likely to be approximately 400 RBPs (Cook et al. 2011; Ray et al. 2013; Dominguez et al. 2018). The major families of RBPs contain canonical RBDs such as the RNA recognition motif (RRM), CCCH zinc finger, K homology (KH) and cold shock domain (CSD). A smaller number of proteins bind RNA using La, HEXIM, PUF, THUMP, YTH, SAM, and TRIM-NHL domains (Ray et al. 2013). In addition, many noncanonical RBPs that do not contain any of the currently known RBDs have been reported to specifically bind to RNA (see, e.g., Gerstberger et al. 2014).

Various methods have been developed to determine the binding positions and specificities of RBPs. Methods that use cross-linking of RNA to proteins followed by immunoprecipitation and then massively parallel sequencing (CLIP-seq or HITS-CLIP) (for review, see Darnell 2010) and PAR-CLIP (Hafner et al. 2010) can determine RNA positions bound by RBPs in vivo, whereas other methods such as SELEX (Tuerk and Gold 1990), RNA Bind-n-Seq (Lambert et al. 2015; Dominguez et al. 2018), and RNAcompete (Ray et al. 2009) can determine motifs bound by RBPs in vitro. Most high-resolution models derived to date have been determined using RNAcompete or RNA Bind-n-Seq. These methods have been used to analyze large numbers of RBPs from multiple species, including the generation of models for a total of 137 human RBPs (Ray et al. 2013; Dominguez et al. 2018).

The cisBP-RNA database (build 0.6) (Ray et al. 2013) currently lists total of 392 high-confidence RBPs in humans but contains high-resolution specificity models for only 100 of them (Ray et al. 2013). The Encyclopedia of DNA Elements (ENCODE) database that contains human RNA Bind-n-Seq data, in turn, has models for 78 RBPs (Dominguez et al. 2018). In addition, a literature curation–based database (Database of RNA-Binding Protein Specificities [RBPDB]) (Cook et al. 2011) contains experimental data for 133 human RBPs but mostly contains individual target sites or consensus sites and only has high-resolution models for 39 RBPs (by high resolution, we refer to models that are derived from quantitative analysis of binding to all short RNA sequences). Thus, despite the central importance of RBPs in fundamental cellular processes, the precise sequence elements bound by most RBPs remain to be determined. To address this problem, we have in this work developed high-throughput RNA-SELEX (HTR-SELEX) and used it to determine binding specificities of human RBPs. Our analysis suggests that many RBPs prefer to bind structured RNA motifs and can associate with several distinct sequences. The distribution of motif matches in the genome indicates that many RBPs have central roles in regulation of RNA metabolism and activity in cells.

Results

Identification of RNA-binding motifs using HTR-SELEX

To identify binding specificities of human RBPs, we established a collection of canonical and noncanonical full-length RBPs and RNA-binding domains, based on the presence of a canonical RBD (from cisBP-RNA database) (Ray et al. 2013). We also included unconventional RBPs that have been reported to bind to RNA but lack canonical RBDs (Gerstberger et al. 2014). Full-length constructs representing 819 putative RBPs were picked from the ORFeome 3.1 and 8.1 collections (Lamesch et al. 2007). In addition, 293 constructs designed to cover all canonical RBDs within 156 human RBPs were synthesized based on Interpro-defined protein domain calls from Ensembl v76. Most RBD constructs contained all RBDs of a given protein with 15 amino acids of flanking sequence (for details, see Supplemental Table S1). For some very large RBPs, constructs were also made that contained only a subset of their RBDs. Taken together, our clone collection covered 942 distinct proteins (Supplemental Table S1). The RBPs were expressed in *Escherichia coli* as fusion proteins with thioredoxin, incorporating an N-terminal hexahistidine and a C-terminal SBP tag (Jolma et al. 2015).

To identify RNA sequences that bind to the proteins, we subjected the proteins to HTR-SELEX (Fig. 1A). In HTR-SELEX, a 40-bp random DNA sequence containing a sample index and 5′ and 3′ primer binding sequences is transcribed into RNA using T7 RNA

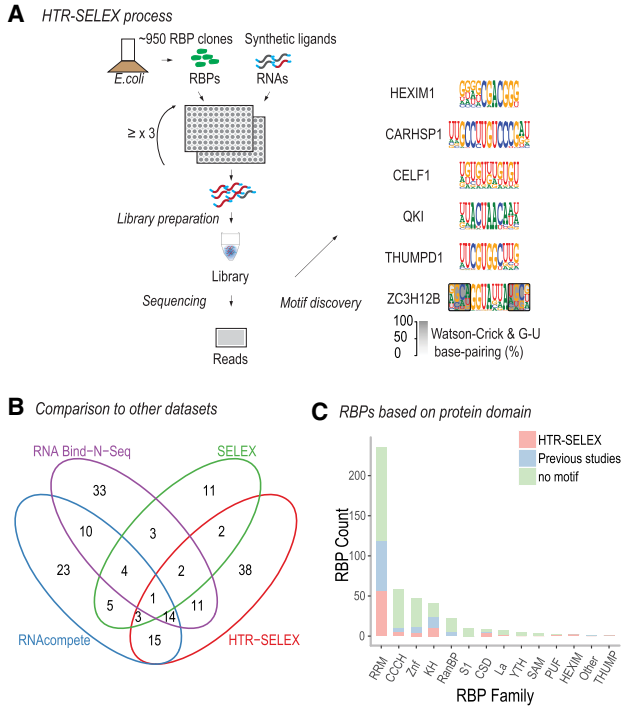

Figure 1. HT RNA-SELEX protocol and data analysis. (*A*) Schematic illustration of the HTR-SELEX process. RBDs or full-length RBPs expressed in *E. coli* as TRX-HIS$_6$-SBP–tagged fusion proteins (*top left*) were purified and incubated with barcoded RNA selection ligands. RNA ligands bound by the proteins were recovered by RT-PCR, followed by in vitro transcription to generate the RNA for the next cycle of SELEX (*middle left*). The procedure was repeated at least three times, and the ligands recovered from the selection cycles were subjected to Illumina sequencing (*bottom left*) with data analysis to generate binding specificity models (*right*). (*B*) Comparison of the number of RBPs with motifs derived in the present study (HTR-SELEX) with the number of RBPs for which motifs were previously derived using RNA Bind-n-Seq (RBNS) (Dominguez et al. 2018), SELEX, and/or RNAcompete (cisBP-RNA version 0.6) (Ray et al. 2013). Note that our analysis revealed motifs for 38 RBPs for which a motif was not previously known. (*C*) Distribution of RBPs with motifs classified by the structural family of their RBDs. RBPs with motifs reported by Ray et al. (2013) and Dominguez et al. (2018) are shown in blue, and RBPs for which motifs were not reported there but determined using HTR-SELEX in this study are in red. RBPs with no motifs are in green.

polymerase and incubated with the individual proteins in the presence of RNase inhibitors, followed by capture of the proteins using metal-affinity resin. After washing and RNA recovery, a DNA primer is annealed to the RNA, followed by amplification of the bound sequences using a reverse-transcription polymerase chain reaction (RT-PCR) using primers that regenerate the T7 RNA polymerase promoter. The entire process is repeated up to a total of four selection cycles. The amplified DNA is then sequenced, followed by identification of motifs using the Autoseed pipeline (Nitta et al. 2015) modified to analyze only the transcribed strand (for details, see Methods). HTR-SELEX uses a selection library with very high sequence complexity, allowing identification of long RNA-binding preferences.

The analysis resulted in generation of 145 binding specificity models for 86 RBPs. Most of the results (66 RBPs) were replicated in a second HTR-SELEX experiment. The success rate of our experiments was ~22% for the canonical RBPs, whereas the fraction of the successful noncanonical RBPs was much lower (~1.3%) (Supplemental Table S1; Supplemental Figs. S16, S17). Comparison of our data with a previous data set generated using RNAcompete (Ray et al. 2013) and RNA Bind-n-Seq (Dominguez et al. 2018) and to older data that has been compiled in the RBPDB-database (Cook et al. 2011) revealed that the specificities were generally consistent with the previous findings (Supplemental Figs. S1, S2). HTR-SELEX resulted in generation of a larger number of motifs than the previous systematic studies, and revealed the specificities of 38 RBPs whose high-resolution specificities were not previously known (Fig. 1B). Median coverage per RBD family was 24% (Fig. 1C). Compared with the motifs from previous studies, the motifs generated with HTR-SELEX were also wider and had a higher information content (Supplemental Fig. S3), most likely because the sequences are selected from a more complex library in HTR-SELEX (see also Yin et al. 2017). The median width and information contents of the models were 10 bases and 10 bits, respectively. To validate the motifs, we evaluated their performance against ENCODE eCLIP data (Supplemental Table S8). This analysis revealed that HTR-SELEX motifs were predictive against in vivo data and that their performance was overall similar to motifs generated using RNAcompete (Ray et al. 2013). The benefit of recovering longer motifs was evident in the analysis of TARDBP, whose HTR-SELEX motif clearly outperformed a shorter RNAcompete motif (Supplemental Fig. S20).

Some RBPs bind to RNA as dimers

Analysis of enriched sequences revealed that 31% of RBPs (27 of 86 with an identified motif) could bind to a site containing a direct repeat of the same sequence (Supplemental Fig. S4; Supplemental Tables S1, S2). Most of these RBPs (15 of 27) had multiple RBDs, which could bind similar sequences, as has been reported previously in the case of ZRANB2 (Loughlin et al. 2009). However, such direct repeats were also bound by RBPs having only a single RBD (12 of 27), suggesting that some RBPs could form homodimers or interact to form a homodimer when bound to RNA (Supplemental Table S2). The gap between the direct repeats was generally short, with a median gap of 5 nucleotides (nt) (Supplemental Fig. S4). To determine whether the gap length preferences identified by HTR-SELEX were also observed in sites bound in vivo, we compared our data against existing in vivo data for four RBPs for which high-quality PAR-CLIP and HITS-CLIP derived data were available from previous studies (Hafner et al. 2010; Farazi et al. 2014; Weyn-Vanhentenryck et al. 2014). We found that preferred spacing identified in HTR-SELEX was in most cases (three out of four) also observed in the in vivo data. However, the gap length distribution observed in vivo extended to longer gaps than that observed in HTR-SELEX (Supplemental Fig. S5), suggesting that such lower-affinity spacings could also have a biological role in RNA folding or function.

Recognition of RNA structures by RBPs

Unlike double-stranded DNA, RNA folds into complex, highly sequence-dependent three-dimensional structures. To analyze whether RBP binding depends on RNA secondary structure, we identified characteristic patterns of dsRNA formation by identifying correlations between all two base positions either within the motif or in its flanking regions, using a measure described by Nitta et al. (2015) that is defined by the difference between the observed count of combinations of a given set of two bases and their expected count based on a model that assumes independence of the positions (Fig. 2A). The vast majority of the observed deviations from the independence assumption were consistent with the formation of an RNA stem–loop structure (example in Fig. 2B; Supplemental Fig. S15). In addition, we identified one RBP, LARP6, that bound to multiple motifs (Supplemental Figs. S6, S19B), including a predicted internal loop embedded in a double-stranded RNA stem (Fig. 2C). This binding specificity is consistent with the earlier observation that LARP6 binds to stem–loops with internal loops found in mRNAs encoding the collagen proteins COL1A1, COL1A2, and COL3A1 (Supplemental Fig. S6; Cai et al. 2010).

In total, 69% (59 of 86) of RBPs recognized linear sequence motifs that did not appear to have a preference for a specific RNA secondary structure. The remaining 31% (27 of 86) of RBPs could bind at least one motif with predicted structure (structured motif hereafter) (Fig. 2D); this group included several known structure-specific RBPs, such as RBFOX1 (Chen et al. 2016), RC3H1, RC3H2 (Leppek et al. 2013), RBMY1E, RBMY1F, RBMY1J (Skrisovska et al. 2007), and HNRNPA1 (Chen et al. 2016; Orenstein et al. 2018). A total of 15 RBPs bound only to structured motifs, whereas 12 RBPs could bind to both structured and unstructured motifs. For example, both linear and structured motifs were detected for RBFOX proteins; binding to both types of motifs was confirmed by analysis of eCLIP data (see Supplemental Fig. S20A).

The median length of the stem region observed in all motifs was 5 bp, and the loops were between 3 and 15 bases long, with a median length of 11 (Fig. 2E). Of the different RBP families, KH and HEXIM proteins only bound linear motifs, whereas proteins from RRM, CSD, zinc finger, and La-domain families could bind to both structured and unstructured motifs (Supplemental Fig. S7).

To model RBP binding to stem–loop structures, we developed a simple stem–loop model (SLM) (Fig. 3; Supplemental Table S2–S4). This model describes the loop as a PWM, and the stem by a nucleotide pair model in which the frequency of each combination of two bases at the paired positions is recorded. In addition, we developed two different visualizations of the model: a T-shaped motif that describes the mononucleotide distribution for the whole model, and the frequency of each set of bases at the paired positions by thickness of edges between the bases (Fig. 3), and a simple shaded PWM in which the stem part is indicated by a gray background where the darkness of the background indicates the fraction of bases that pair with each other using Watson–Crick or G:U base pairs (Fig. 3). Analysis of the SLMs for each structured motif

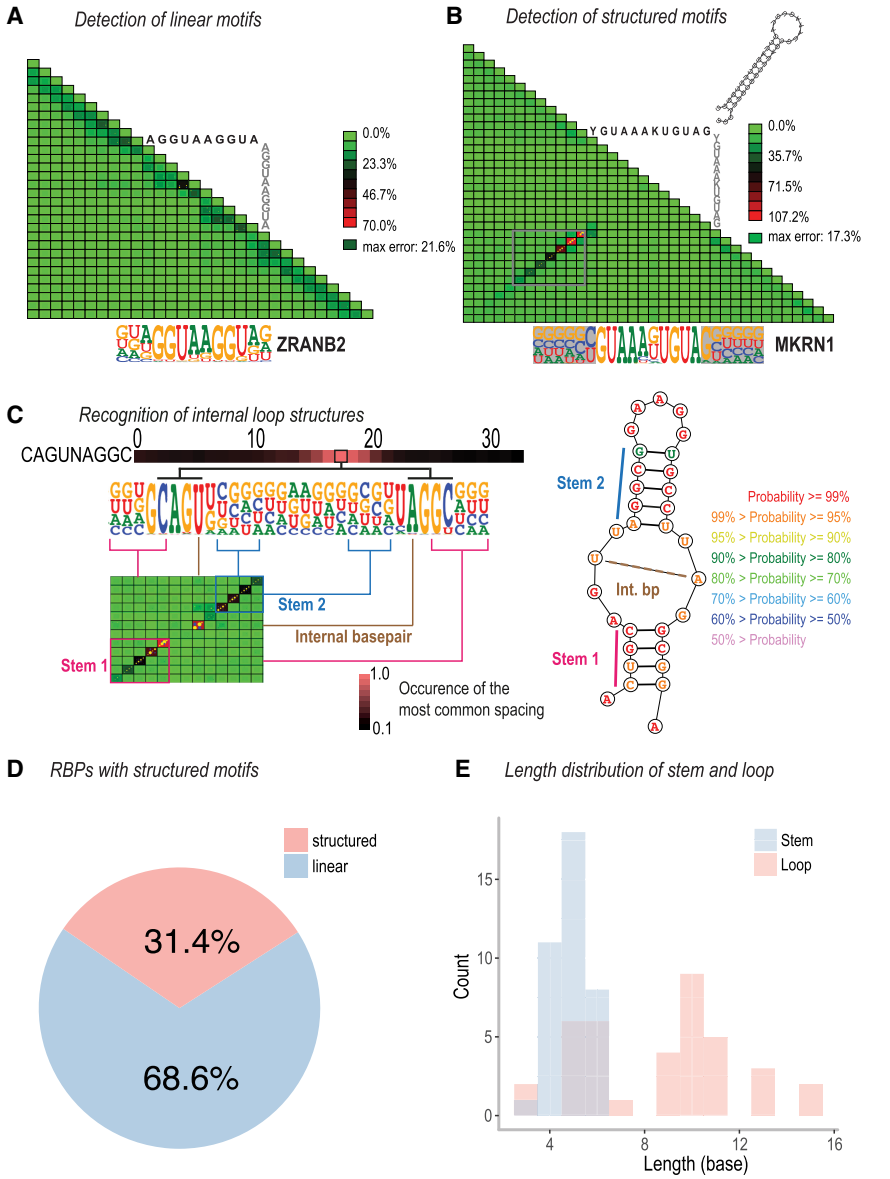

Figure 2. Detection of linear or structured RNA-binding models. (*A*) ZRANB2 binds to a linear RNA motif. The motif of ZRANB2 and the seed used to derive it are shown *below* and *above* the triangular correlation heat map, respectively. The heat map illustrates deviation of the observed nucleotide distributions from those predicted by a mononucleotide model in which bases are independent. (*B*) MKRN1 binds preferentially to a stem–loop. Note a diagonal series of red tiles (boxed) that indicates pairs of bases whose distribution deviates from the independence assumption. These bases are shaded in the motif below the triangle. The interdependency occurs between bases that are at the same distance from the center of the motif, consistent with formation of a stem–loop structure. *Right top*: An RNAfold-predicted stem–loop structure for a sequence that was highly enriched in the experiment. (*C*) LARP6 binds to a complex internal loop RNA structure. The *left* panel indicates the dinucleotide dependencies with the heat map on *top* representing the preferred spacing length between base-pairing sequences of stem 1, whereas the *right* panel presents a predicted structure of the bound RNA. The dashed line in the structure denotes the internal base pair. (*D*) Fraction of RBPs with linear and structured binding specificities. RBPs with at least one structured specificity are counted as structured. (*E*) Length distribution of stem and loop for the structured motifs.

indicated that on average, the SLM increased the information content of the motifs by 4.2 bits (Supplemental Fig. S8). Independent secondary structure analysis performed using RNAfold indicated that as expected from the SLM, >80% of individual sequence reads for MKRN1 had more than four paired bases, compared with ~15% for the control RBP (ZRANB2) for which a structured motif was not identified (Fig. 3).

Classification of RBP motifs

To analyze the motif collection globally, we developed PWM and SLM models for all RBPs. To compare the motifs, we determined their similarity using SSTAT (Pape et al. 2008). To simplify the analysis, PWM models were used for this comparison even for RBPs that bound to the structured motifs. We then used the dominating set method (Jolma et al. 2013) to identify a representative set of distinct motifs (Supplemental Fig. S9). Comparison of the motifs revealed that in general, the specificities of evolutionarily related RBPs were similar (Fig. 4A–F; Supplemental Fig. S9). For the largest family, RRM, a total of 96 motifs were represented by 47 specificity classes, whereas the smaller families CCCH, KH, CSD, and HEXIM were represented by nine, 10, six, and one classes, representing 17, 11, seven, and two individual motifs, respectively (Supplemental Fig. S9).

Analysis of the dinucleotide content of all motifs revealed unexpected differences in occurrence of distinct dinucleotides within the PWMs. The dinucleotides GG, GU, UG, and UU were much more common than other dinucleotides (fold change 2.75; $P<0.00225$; *t*-test) (Fig. 4G). This suggests that G and U bases are most commonly bound by RBPs. This effect could be in part because of structural motifs, in which G and U can form two different base-pairs. Furthermore, many RBPs function in splicing, and their motifs preferentially match sequences related to the G-U–rich splice donor sequence A/UG:GU (Supplemental Data S1–S4). However, G and U enrichment cannot be explained by structure alone, as the unstructured motifs were also enriched in G and U. One possibility is that the masking of G and U bases by protein binding may assist in folding of RNA to defined structures, as G and U bases have lower specificity in base-pairing than C and A, owing to the presence of the non-Watson–Crick G:U base pairs in RNA. The enrichment of G and U bases in RBP motifs was also previously reported in a different motif set discovered using a different method, RNA Bind-n-Seq (Dominguez et al. 2018). (For comparison with RNAcompete, see Supplemental Fig. S21.)

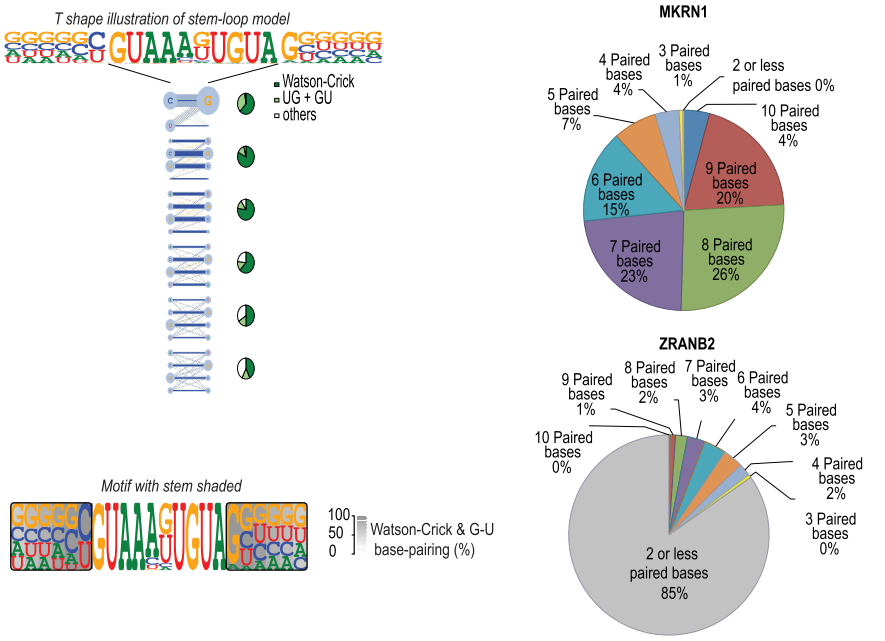

Figure 3. Comparison between linear PWM and stem–loop (SLM) models. (*Left*) Visualization of the stem–loop models. (*Top*) A T-shape model shows a horizontal loop and a vertical stem where the frequency of each base combination is shown. Bases are aligned so that Watson–Crick base pairs orient horizontally. Pie-charts show frequency of Watson–Crick (green) and G-U base pairs (light green) compared with other pairs (gray) that do not form canonical dsRNA base pairs at each position of the predicted stem. (*Bottom*) A linear visualization in which the base-pairing frequency is indicated by the darkness of gray shading is also shown. (*Right*) RNA secondary structure prediction analysis using RNAfold reveals that sequences flanking MKRN1 loop sequence form base pairs (*top*), whereas bases on the flanks of ZRANB2 matches (*bottom*) are mostly unpaired.

Most RBPs bound to only one motif. However, 41 RBPs could bind to multiple different motifs, which present a limited degree of similarity, generally reflecting previous observations that many RBPs are relatively promiscuous in their motif recognition (Fig. 5; Draper 1999; Jones et al. 2001; Mackereth and Sattler 2012). Of these, 19 had multiple RBDs that could explain the multiple specificity. However, 22 RBPs could bind to multiple motifs despite having only one RBD, indicating that individual RBPs are commonly able to bind to multiple RNA sequences. In five cases, the differences between the primary and secondary motif could be explained by a difference in spacing between the two half-sites. In 12 cases, one of the motifs was structured and the other linear. In addition, in eight RBPs the primary and secondary motifs represented two different structured motifs, in which the loop length or the loop sequence varied (Fig. 5). In addition, for four RBPs, we recovered more than two different motifs. The most complex binding specificity we identified belonged to LARP6 (Fig. 5; Supplemental Fig. S10), which could bind to multiple simple linear motifs, multiple dimeric motifs, and the internal loop-structure described above.

Conservation and occurrence of motif matches

We next analyzed the enrichment of the motif occurrences in different classes of human transcripts. The normalized density of motif matches for each RBP at both strands of DNA was evaluated relative to the following features: transcription start sites (TSSs), splice donor and acceptor sites, and translational start and stop positions (see Supplemental Fig. S11; for full data, see Supplemental Data S1–S4). This analysis revealed that many RBP recognition motifs were enriched at splice junctions. The most enriched linear motif in splice donor sites belonged to ZRANB2, a known regulator of alternative splicing (Fig. 6A; Loughlin et al. 2009). Analysis of matches to structured motifs revealed even stronger enrichment of motifs for ZC3H12A, -B, and -C to splice donor sites (Fig. 6A). These results suggest a novel role for ZC3H12 proteins in regulation of splicing. The motifs for both ZRANB2 and ZC3H12 protein factors were similar but not identical to the canonical splice donor consensus sequence ag|GU[g/a]agu (Fig. 6A) that is recognized by the spliceosome, suggesting that these proteins may act by binding to a subset of splice donor sites.

Analysis of splice acceptor sites also revealed that motifs for known components of the spliceosome, such as RBM28 (Damianov et al. 2006), were enriched in introns and depleted in exons (Supplemental Data S1–S4). Several motifs were also enriched at the splice junction, including the known regulators of splicing IGF2BP1 and ZFR (Supplemental Data S1–S4; Haque et al. 2018; Huang et al. 2018). In addition, we found several motifs that mapped to the 5′ of the splice junction, including some known splicing factors such as QKI (Hayakawa-Yano et al. 2017) and ELAVL1 (Bakheet et al. 2018), as well as some factors such as DAZL, CELF1 and BOLL for which a role in splicing has to our knowledge not been reported (Fig. 6A; Supplemental Data S1–S4; Rosario et al. 2017; Xia et al. 2017).

To determine whether the identified binding motifs for RBPs are biologically important, we analyzed the conservation of the motif matches in mammalian genomic sequences close to splice junctions. This analysis revealed strong conservation of several classes of motifs in the transcripts (Fig. 6B; Supplemental Table S6), indicating that many of the genomic sequences matching the motifs are under purifying selection.

Matches to both ZRANB2 and ZC3H12 motifs were enriched in 5′ regions of the sense strands of known transcripts, but not on the corresponding antisense strands. However, no enrichment was detected in the potential transcripts that would originate from the same promoters and extend in a direction opposite to that of the mRNAs (Fig. 6C). These results suggest that ZRANB2 and ZC3H12 motifs could have a role in differentiating between forward and reverse strand transcripts that originate from bidirectional promoters.

We also used Gene Ontology enrichment analysis to identify motifs that were enriched in specific types of mRNAs. This analysis revealed that many RBP motifs are specifically enriched in particular classes of transcripts. For example, we found that MEX3B motifs were enriched in genes involved in the type I interferon-mediated signaling pathway (Fig. 6D; Supplemental Table S7).

Taken together, our analysis indicates that RBP motifs are biologically relevant, as matches to the motifs are conserved and occur specifically in genomic features and in transcripts having specific biological roles.

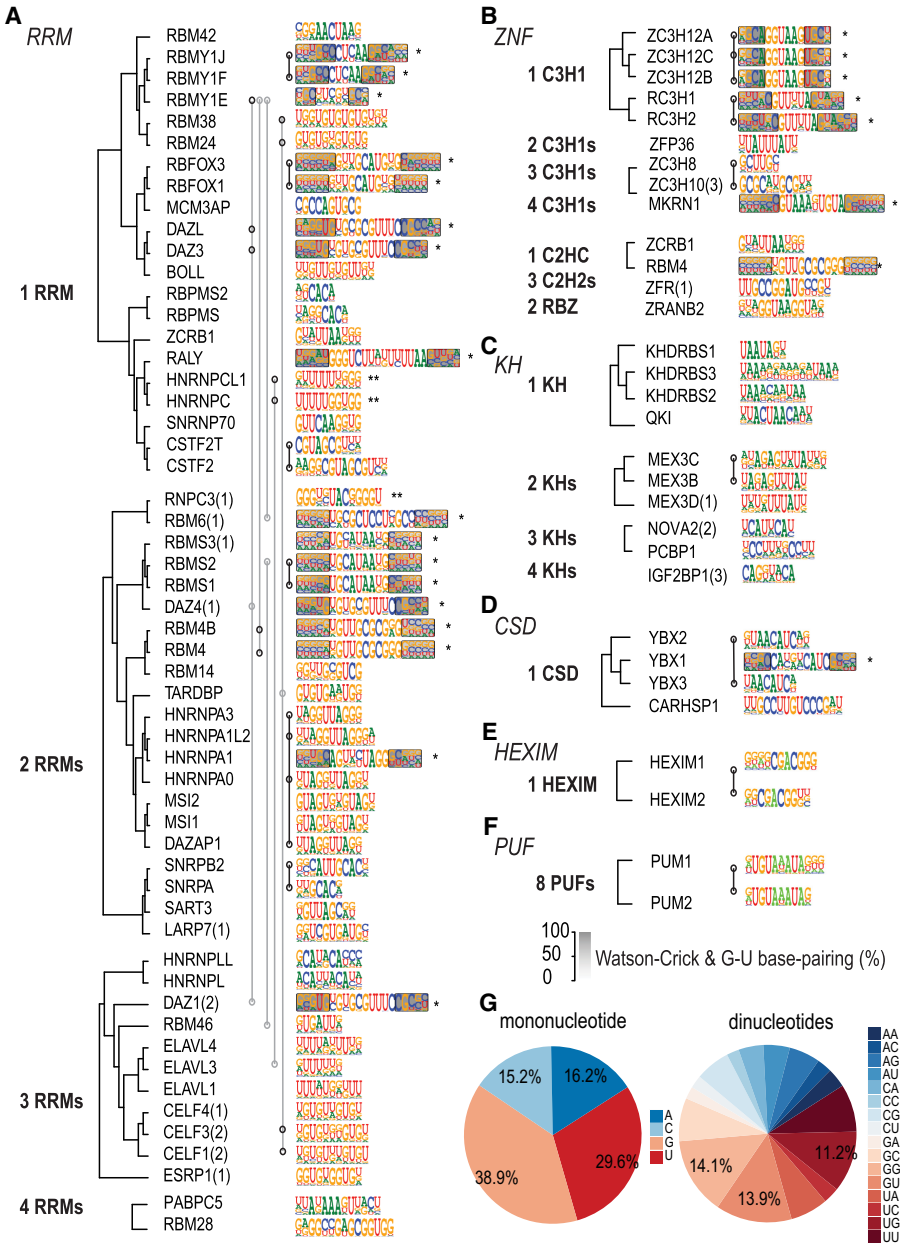

Figure 4. Comparison between the HTR-SELEX motifs. (*A–F*) Similar RBPs bind to similar motifs. Motifs were classified into six major categories based on structural class of the RBPs. Dendrograms are based on amino acid alignment using PRANK (Löytynoja and Goldman 2005). Within the RRM family, RBPs with different numbers of RRMs were grouped and aligned separately; if fewer RBDs were included in the construct used, the number of RBDs is indicated in parentheses (see also Supplemental Table S1). Motifs shown are the primary motif for each RBP. Asterisks indicate a stem–loop structured motif, with the gray shading showing the strength of the base-pairing at the corresponding position. Two asterisks indicate that the RBP can bind to a structured secondary motif. Motifs that are similar to each other based on SSTAT analysis (covariance threshold 5×10^{-6}) are indicated by open circles connected by lines. Only families with more than one representative HTR-SELEX motif are shown. (*G*) RBPs commonly prefer sequences with G or U nucleotides. Frequencies of all mononucleotides (*left*) and dinucleotides (*right*) across all of the RBP motifs. Note that G and U are overrepresented.

Structural analysis of ZC3H12B bound to RNA

The ability of the cytoplasmic ZC3H12 proteins to bind to splice donor-like sequences suggests that these proteins may be involved in recognition of unspliced cellular mRNA or viral transcripts in the cytoplasm, both of which would be subject to degradation. Indeed, the ZC3H12 proteins, which are conserved across metazoa, have been linked to protective responses against viral infection (Fu and Blackshear 2017; Wilamowski et al. 2018). Moreover, these proteins (and all our constructs) contain both C3H1 RBD and a PIN RNase domain, and previous studies have indicated that the ZC3H12 proteins are RNA endonucleases that rapidly degrade specific RNAs (Wilamowski et al. 2018).

To further explore our unexpected finding that these proteins are stably associated with splice donor-like sequences, we solved the structure of ZC3H12B together with a 21-base RNA sequence enriched in HTR-SELEX at 3.3 Å resolution (Fig. 7A,B; for statistics of data collection and refinement, see Supplemental Table S5). Unexpectedly, we found that the RNA was bound to the PIN nuclease domain and not to the conventional RNA-binding domain (C3H1), which was not resolved in our structure. As reported previously for ZC3H12A (Xu et al. 2012), in the structure, ZC3H12B appears as a dimeric protein, with single Mg^{2+} ion coordinated at each active site. The dimer is held together by a relatively large contact surface (1008.2 $Å^2$); however, it is predicted to exist as a monomer in solution (complex significance score CSS = 0) (see also Xu et al. 2012). Similarly, the other contacts observed in the asymmetric unit of the crystal, including the RNA–RNA contact (877.0 $Å^2$) and protein dimer-to-dimer contact (1028.1 $Å^2$), appear too weak to exist in solution (CSS = 0 for both). Taken together, despite the 2:1 protein–RNA stoichiometry in the crystals, it is likely that the complex exists as either 2:2 or 1:1 in solution. However, confirmation of this prediction awaits further biophysical analysis.

In our structure, only one of the active sites is occupied by RNA; the protein–RNA interaction is predicted to be stable (CSS ≈ 0.6). The overall structure of the RNA-bound ZC3H12B PIN domain is highly similar to the unbound domain and to the previously reported structure of the free PIN domain of ZC3H12A (Supplemental Fig. S12). The active site is relatively shallow, and the magnesium is coordinated by only one direct amino-acid contact (Asp280) together with five water molecules.

In the structure, the segment of the RNA backbone bound to the active site adopts a specific horseshoe-like shape that is highly similar to an inhibitory RNA bound to an unrelated RNase DIS3 (Supplemental Fig. S13; Weick et al. 2018) in the structure of the human exosome (PDB: 6D6Q). The protein binds to five RNA

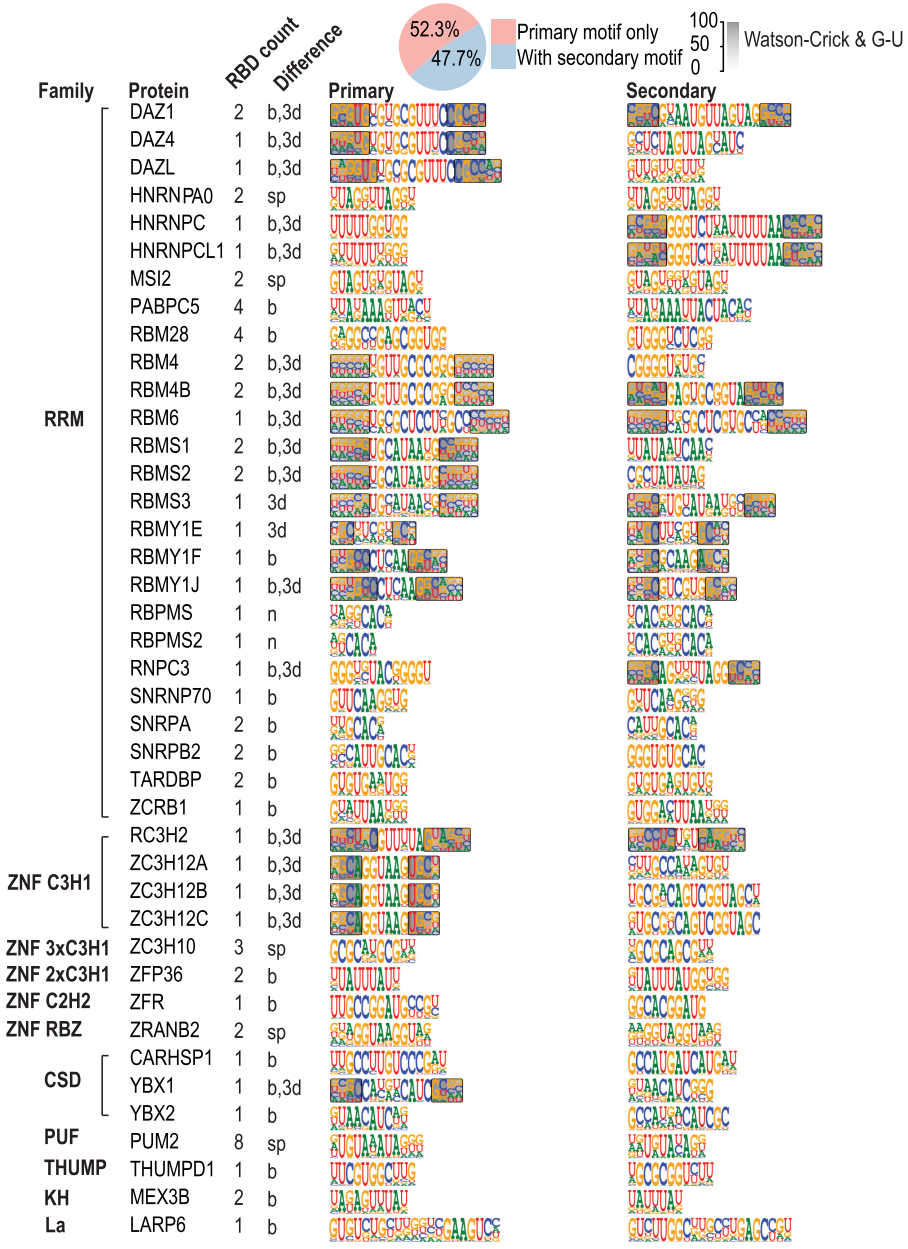

Figure 5. Many RBPs can recognize more than one motif. (*Top*) Pie chart indicates fraction of RBPs that recognize more than one motif. Primary (*left*) and secondary (*right*) motifs are shown, classified according to the RBP structural family. Number next to the RBD name indicates the number of RBDs in the construct used, and the letters indicate how the two motifs are different from each other, as follows: difference in number of half-sites (n), half-site spacing (sp), base recognition (b), and/or secondary structure (3d).

bases, consistently with earlier observations suggesting that a minimum length of RNA is needed for the endonuclease activity (Lin et al. 2013). The RNA is bound mainly via interactions to the phosphate backbone and ribose oxygens; only G13 and G14 are recognized by direct hydrogen bonds between Asp244 and N3 of the guanine G13 and N3 of G14. G17, in turn, is recognized by a hydrogen bond between Arg301 and O2′ of the ribose and a water-mediated hydrogen bond between Asp302 and O6 of the guanine (Fig. 7C–F; Supplemental Fig. S18). The specificity toward the central GUA trinucleotide that is common to most motifs bound by the ZC3H12 family (Fig. 7A) is most likely determined by an extensive water network connected to the magnesium ion and by hydrogen bonding to the symmetric molecule of RNA (G14 to U11, U15 to A9, A16 to G6) (Supplemental Fig. S14).

The structure suggests that the RNA molecule bound to the PIN domain is a relatively poor substrate to the RNase, as although the RNA backbone is tightly bound and oriented toward the active site, the phosphate between U15 and A16 remains still relatively far from the magnesium ion.

Discussion

In this work, we have determined the RNA-binding specificities of a large collection of human RBPs. The tested proteins included both proteins with canonical RNA-binding domains and putative RBPs identified experimentally (Ray et al. 2013; Gerstberger et al. 2014). The method used for analysis involved selection of RNA ligands from a collection of random 40-nt sequences. Compared with previous analyses of RBPs, the HTR-SELEX method allows identification of predicted structured motifs and of motifs that are relatively high in information content. The method can identify simple sequence motifs or structured RNAs, provided that their information content is less than ~40 bits. However, because of the limit on information content and the requirement of relatively high-affinity binding, the method does not generally identify highly structured RNAs that in principle could bind to almost any protein. Consistent with this, most binding models that we could identify were for proteins containing canonical RBPs.

Motifs were identified for a total of 86 RBPs. A large fraction of all RBPs (47%) could bind to multiple distinctly different motifs. The fraction is much higher than that observed for double-stranded DNA-binding transcription factors, suggesting that sequence recognition and/or individual binding domain arrangement on single-stranded RNA can be more flexible than on dsDNA (see Draper 1999; Jones et al. 2001; Mackereth and Sattler 2012). Analysis of the mononucleotide content of all the models also revealed a bias toward recognition of G and U over C and A (see also Dominguez et al. 2018). This may reflect the fact that the formation of RNA structures is largely based on base-pairing and that G and U are less specific in their base-pairings that C and A. Thus, RBPs that mask G and U bases increase the overall specificity of RNA folding in cells.

Similar to proteins, depending on sequence, single-stranded nucleic acids may fold into complex and stable structures or

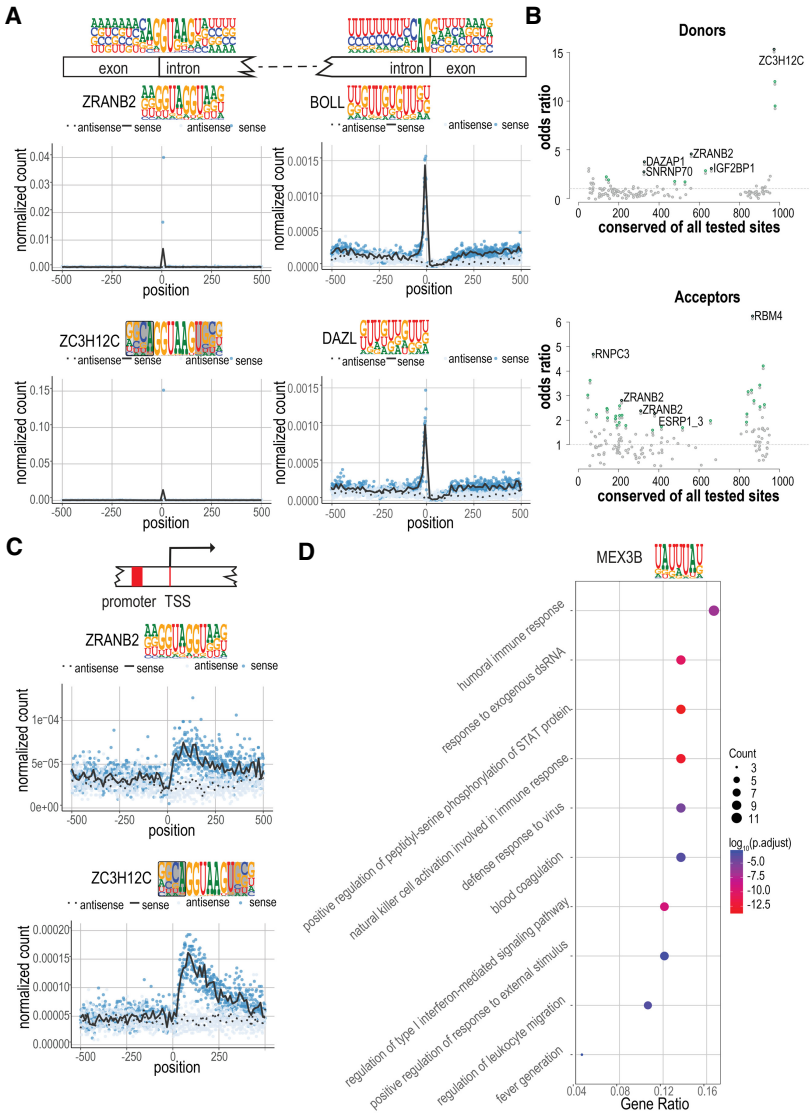

Figure 6. RBP motif matches are conserved and enriched in distinct sequence features and classes of transcripts. (*A*) Strong enrichment of RBP motif matches at or near the splicing donor and acceptor sites. Mononucleotide frequencies at splice donor and acceptor sites are shown on *top*, above the gene schematic. (*Left*) Meta-plots indicate the enrichment of ZRANB2 and ZC3H12C motif matches at splice donor sites. (*Right*) Enrichment of BOLL and DAZL at splice acceptor sites. Blue dots indicate the number of matches in the sense strand at each base position; black line indicates the locally weighted smoothing (LOESS) curve in 10-base sliding windows. Corresponding values for the antisense strand are shown as light blue dots and dotted black line, respectively. (*B*) The conservation of motif matches in sense versus antisense strand. Odds ratio of preferential conservation of a match in the sense strand (*y*-axis) is shown as a function of the total number of conserved motif matches (*x*-axis) (for details, see Methods). Motifs for which conservation is significantly associated with sense strand (one-sided Fisher's exact test) are shown in green. The five motifs with the smallest *P*-values are indicated in black and named. (*C*) Enrichment of ZRANB2 and ZC3H12C motif matches near transcription start sites (TSSs). Note that matches are only enriched on the sense strand downstream from the TSS. (*D*) Gene Ontology (GO) enrichment of MEX3B motif matches. The top 100 genes with highest motif-matching score density were used to conduct the GO enrichment analysis. The enriched GO terms were simplified by their similarity (cutoff = 0.5). The fraction of genes and their counts in the GO categories are also shown (gene ratio, count, respectively).

remain largely disordered. Most RBPs preferred short linear RNA motifs, suggesting that they recognize RNA motifs found in unstructured or single-stranded regions. However, ~31% of all RBPs preferred at least one structured motif. The vast majority of the structures that they recognized were simple stem–loops, with relatively short stems, and loops of 3 to 15 bases. Most of the base specificity of the motifs was found in the loop region, with only one or few positions in the stem displaying specificity beyond that caused by the paired bases. This is consistent with the structure of fully paired double-stranded RNA in which base-pair edge hydrogen-bonding information is largely inaccessible in the deep and narrow major groove. In addition, we identified one RBP that bound to a more complex structure. LARP6, which has previously been shown to bind to RNA using multiple RBPs (Martino et al. 2015), recognized an internal loop structure in which two base-paired regions were linked by an uneven number of unpaired bases.

Compared with TFs, which display complex dimerization patterns when bound to DNA, RBPs displayed simpler dimer spacing patterns. This is likely because the backbone of a single-stranded nucleic acid has rotatable bonds. Thus, cooperativity between two RBDs requires that they bind to relatively closely spaced motifs.

Analysis of in vivo–bound sequences revealed that the HTR-SELEX motifs were predictive of binding inside cells as determined by eCLIP. However, it is expected that similarly to the case of DNA-bound transcription factors, all strong motif matches will not be occupied in vivo. This is because binding in vivo will depend on competition between RBPs, their localization, and the secondary structure of the full RNAs. Analysis of the biological roles of the RBP motif matches further indicated that many motif matches were conserved and specifically located at genomic features such as splice junctions. In particular, our analysis suggested a new role for ZC3H12, BOLL, and DAZL proteins in regulating alternative splicing and for MEX3B in binding to type I interferon-regulated genes. In particular, the binding of the antiviral cytoplasmic ZC3H12 proteins (Lin et al. 2013; Habacher and Ciosk 2017) to splice junctions may have a role in their antiviral activity, as endogenous cytoplasmic mRNAs are depleted of splice donor sequences. As a large number of novel motifs were generated in the study, we expect that many other RBPs will have specific roles in particular biological functions.

Although we included the ZC3H12 proteins to our study because they contained the known, canonical RNA-binding domain C3H1, our structural analysis revealed that the RNA was

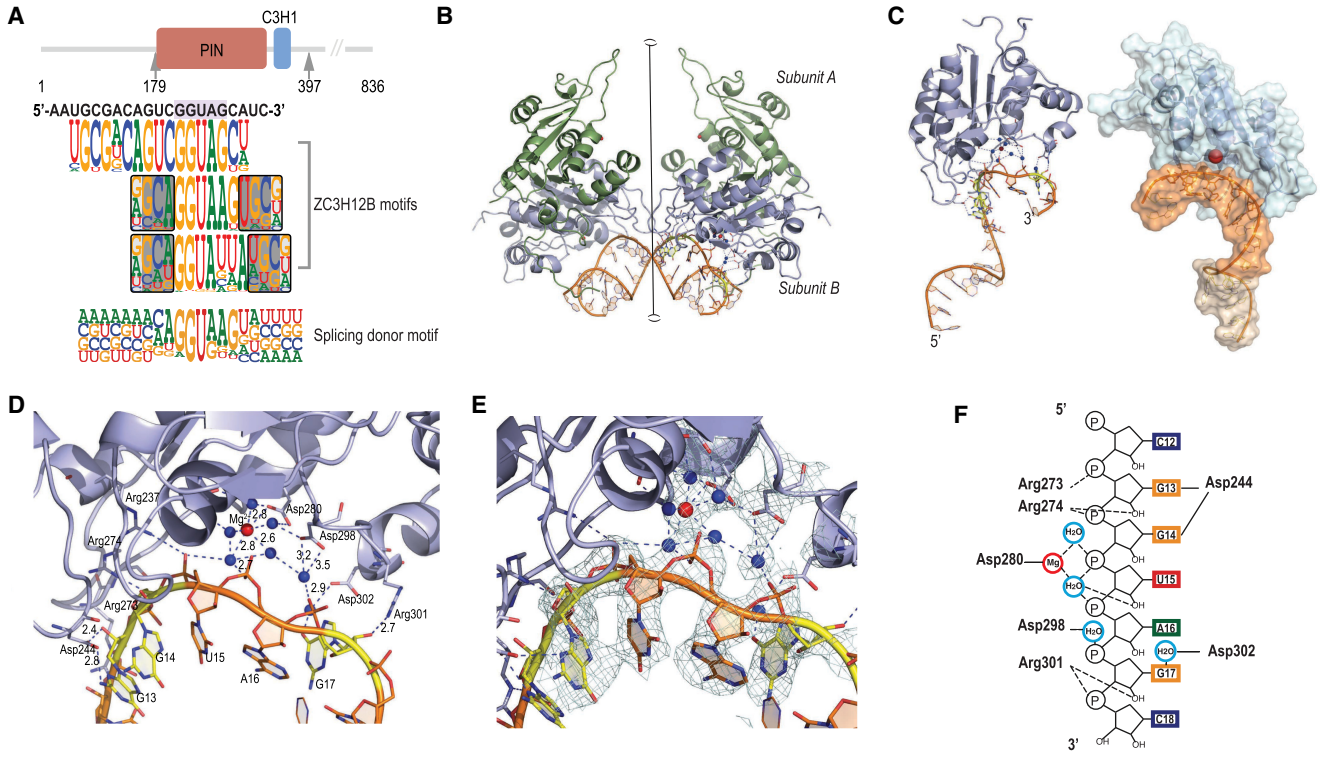

Figure 7. Structural basis of RNA motif recognition by ZC3H12B. (*A*) Schematic representation of the domain structure of ZC3H12B. The arrows indicate the first and the last amino acid of the construct used for crystallization, containing both the PIN domain (residues 181–350) (Senissar et al. 2017) and the known RNA-binding C3H1 zinc finger domain (residues 355–380) (Lai et al. 2002; Hudson et al. 2004). RNA sequence used for crystallization and all ZC3H12B motifs and for the splice donor motif are shown *below* the cartoon. Note that all these motifs contain the sequence GGUA. (*B*) Figure shows two asymmetric units of the crystals of RNA-bound ZC3H12B (a total of two RNAs, four ZC3H12B proteins). Only the PIN domain is visible in the structure. The crystals belong to the $P4_32_12$ space group, and the asymmetric unit contains one protein dimer of two identical monomers presented in green (subunit A) and blue (subunit B) and one RNA molecule. This dimer is similar to the dimer found in the structure of ZC3H12A (PDB: 3V33) (Xu et al. 2012). Note that the contact between the two dimers of ZC3H12B around the twofold crystallographic axis (vertical line) is primarily mediated by the two RNA chains. Red and blue spheres represent Mg^{2+} ions and water molecules, respectively. For clarity, only the water molecules found in the active site are shown. Dashed lines represent hydrogen bonds (*right* side). The residues involved in the protein–RNA contacts are shown as ball-and-stick models, and the nucleotides involved in hydrogen bonds with these residues are in yellow. Notice that only the active site of subunit B of the AB dimer is occupied by an RNA molecule. (*C*) The structure of ZC3H12B PIN domain. (*Left*) The PIN domain is composed of a central beta-sheet surrounded by alpha-helices from both sides. The RNA molecule is bound near the Mg^{2+} ion by the -GGUAG- sequence, which is located close to the 3′ end of the cocrystallized RNA. (*Right*) Surface model shows the shape of the active site bound by RNA (brown), with the weakly coordinated Mg^{2+} ion. Waters are omitted for clarity. Note the horseshoe-like shape of the RNA backbone at the active site (orange). (*D*) A closeup image of the RNA fragment bound to the catalytic site of ZC3H12B. Mg^{2+} ion is shown as a red sphere; the water molecules are represented as blue spheres, with dashed lines representing hydrogen bonds. Note that phosphates of U15, A16, and G17 interact with the Mg^{2+} ion via water molecules. The Mg^{2+} ion is coordinated by five water molecules that also mediate contact with one of the side-chain oxygen atoms of Asp280 as well as Asp195 and Asp298 and phosphate groups of RNA. Thus, the octahedral coordination of the Mg^{2+} ion is distorted, and the ion is shifted from the protein molecule toward the RNA chain, interacting with the RNA via an extensive network of hydrogen bonds. The RNA backbone is slightly bent away from the protein, suggesting that the sequence is a relatively poor substrate. The presence of only one magnesium ion and the positions of water molecules correspond to the cleavage mechanism suggested for the HIV-1 RNase H (Keck et al. 1998). (*E*) The image in *D* annotated with the 2Fo–Fc electron density map contoured at 1.5 σ (light green mesh). (*F*) Schematic representation of interactions between protein, the Mg^{2+} ion, and RNA. Solid lines represent contacts with RNA bases, whereas hydrogen bonds to ribose and phosphates are shown as dashed lines. Nucleotide bases are presented as rectangles and colored as follows: G, yellow; A, green; U, red; and C, blue. Water molecules and Mg^{2+} ion are shown as light blue and red rings, respectively.

instead recognized specifically by the PIN domain, which has not been previously linked to sequence-specific recognition of RNA. The PIN domain active site is relatively shallow and contains one weakly coordinated magnesium ion. The active site was occupied by the RNA motif sequence that adopted a very specific horse-shoe-like shape. The bound RNA is most likely a poor substrate for the RNase, but further experiments are needed to establish the binding affinity of and enzymatic parameters for the bound RNA species. Its binding mechanism, however, suggests that proteins containing small molecule binding pockets or active sites can bind to relatively short, structured RNA molecules that insert into the pocket. This finding indicates that it is likely that all human proteins that bind sequence specifically to RNA motifs have not yet been annotated. In particular, several recent studies have found that many cellular enzymes bind to RNA (Hentze et al. 2018; Queiroz et al. 2019). The structure of ZC3H12B bound to RNA may thus also be important in understanding the general principles of RNA recognition by such unconventional RBPs (Hentze et al. 2018; Queiroz et al. 2019).

Our results represent the largest single systematic study of human RBPs to date. This class of proteins is known to have major roles in RNA metabolism, splicing, and gene expression. However, the precise roles of RBPs in these biological processes are poorly understood, and in general, the field has been severely understudied. The generated resource will greatly facilitate research in this important area.

Methods

Clone collection, protein expression, and structural analysis

Clones were either collected from the human ORFeome 3.1 and 8.1 clone libraries (full-length clones) or ordered as synthetic genes from GenScript (RBP constructs). As in our previous work (Jolma et al. 2013), protein-coding synthetic genes or full-length ORFs were cloned into pETG20A-SBP to create an *E. coli* expression vector that allows the RBP or RBD cDNAs to be fused N-terminally to thioredoxin+6XHis and C-terminally to SBP-tags. Fusion proteins were then expressed in the Rosetta P3 DE LysS *E. coli* strain (Novagen) using an autoinduction protocol (Jolma et al. 2015). For protein purification and structural analysis using X-ray crystallography, see the Supplemental Methods.

HTR-SELEX assay

The HTR-SELEX assay was performed in 96-well plates, where each well contained an RNA ligand with a distinct barcode sequence. A total of three or four cycles of the selection reaction was then performed to obtain RNA sequences that bind to the RBPs. Selection reactions were performed as follows: ~200 ng of RBP was mixed on ice with ~1 µg of the RNA selection ligands to yield an approximate 1:5 molar ratio of protein to ligand in 20 µL of Promega buffer (50 mM NaCl, 1 mM $MgCl_2$, 0.5 mM Na_2EDTA, and 4% glycerol in 50 mM Tris-Cl at pH 7.5). The complexity of the initial DNA library is approximately 10^{12} DNA molecules with 40-bp random sequence (about 20 molecules of each 20-bp sequence on the top strand). The upper limit of detection of sequence features of HTR-SELEX is thus ~40 bits of information content.

The reaction was incubated for 15 min at +37°C followed by additional 15 min at room temperature in 96-well plates (4-titude), after which the reaction was combined with 50 µL of 1:50 diluted paramagnetic HIS-tag beads (His Mag Sepharose excel, GE-Healthcare) that had been blocked and equilibrated into the binding buffer supplemented with 0.1% Tween 20 and 0.1 µg/µL of BSA (molecular biology grade, NEB). Protein–RNA complexes were then incubated with the magnetic beads on a shaker for further 2 h, after which the unbound ligands were separated from the bound beads through washing with a BioTek 405CW plate washer fitted with a magnetic platform. After the washes, the beads were suspended in heat elution buffer (0.5 µM RT-primer, 1 mM EDTA, and 0.1% Tween20 in 10 mM Tris-Cl buffer at pH 7) and heated for 5 min at 70°C followed by cooling on ice to denature the proteins and anneal the reverse transcription primer to the recovered RNA library, followed by reverse transcription and PCR amplification of the ligands using primers that regenerate the T7 promoter sequences. The efficiency of the selection process was evaluated by running a qPCR reaction in parallel with the standard PCR reaction.

PCR products from RNA libraries (indexed by barcodes) were pooled together, purified using a PCR-purification kit (Qiagen), and sequenced using Illumina HiSeq 2000 (55-bp single reads). Data was de-multiplexed and initial data analysis performed using the Autoseed algorithm (Nitta et al. 2015) that was further adapted to RNA analysis by taking into account only the transcribed strand and designating uracil rather than thymine (for detailed description, see the Supplemental Methods).

Comparison of motifs and analysis of their biological functions

To assess the similarity between publicly available motifs and our HTR-SELEX data, we aligned the motifs as described in Supplemental Figure S1. (Jolma et al. 2015). The alignment score for the best alignment was calculated as follows: Max (information content for PWM1 position n, information content for PWM2 position m) × (Manhattan distance between base frequencies of PWM1 position n and PWM2 position m). In regions where there was no overlap, the positions were compared to an equal frequency of all bases. The package SSTAT (Pape et al. 2008) was used to measure the similarity of the RBP PWM motifs, and the dominating set of representative motifs (see Jolma et al. 2013) was generated using a covariance threshold of 5×10^{-6}.

To gain insight into the function of the RBPs, we mapped each motif to the whole human genome (hg38). We applied different strategies for the linear and the stem–loop motifs. For the linear motifs, we identified the motif matches with MOODS (Korhonen et al. 2017) with the following parameter setting: --best-hits 300000 --no-snps. For the stem–loop motifs, we implemented a novel method to score sequences against the SLMs (Supplemental Fig. S19A; see Data access).

We identified the 300,000 best-scored matches in the genome and further included any matches that had the same score as the match with the lowest score, leading to at least 300,000 matches for each motif. As the RNAs analyzed only cover 33% of the genome, this yields approximately 100,000 matches per transcriptome. The constant number of motif matches was used to make comparisons between the motifs simpler. Because of differences in biological roles of the RBPs, further analysis using distinct thresholds for particular RBPs is expected to be more sensitive and more suitable for identifying particular biological features.

The matches were then intersected with the annotated features from the Ensembl database (hg38, version 91), including the splicing donor (DONOR), splicing acceptor (ACCEPTOR), the translation start codon (STARTcodon), the translation stop codon (STOPcodon), and the transcription starting site (TSS). The above features were filtered in order to remove short introns (<50 bp) and features with nonintact or noncanonical start codon or stop codon. The filtered features were further extended 1 kb both upstream and downstream in order to place the feature in the center of all the intervals. The motif matches overlapping the features were counted using BEDTools (version 2.15.0) (Quinlan and Hall 2010) and normalized by the total number of genomic matches for the corresponding motif. For analysis of conservation of motif matches, mutual information analysis, and Gene Ontology enrichment, see the Supplemental Methods.

Data access

The massively parallel sequencing data generated in this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/) database under accession number PRJEB25907. The diffraction data and the model of the ZC3 H12B:RNA complex have been deposited at Protein Data Bank (PDB; https://www.wwpdb.org) under accession code 6SJD. All computer code and scripts developed for this study are available on the GitHub repository (https://github.com/zhjilin/rmap) and as Supplemental Code. Requests for materials should be addressed to J.T. (ajt208@cam.ac.uk).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Drs. Minna Taipale and Bernhard Schmierer for the critical review of the manuscript, as well as Sandra Augsten, Lijuan Hu, and Anna Zetterlund for the technical assistance. The work was

supported by a travel and project grant support (E. Mondragón, L.J.M.) from the Mayo Clinic–Karolinska Institutet collaboration partnership, as well as the Knut and Alice Wallenberg Foundation (KAW 2013.0088) and the Swedish Research Council (postdoctoral grant 2016-00158).

Author contributions: J.T., A.J., and L.J.M. designed the experiments; A.J., E. Mondragón, and Y.Y. performed the SELEX experiments; A.J., J.Z., J.T., K.L, T.K., T.R.H., Q.M., and F.Z. analyzed the data; E. Morgunova and G.B. solved the structure; J.T., A.J., and J.Z. wrote the manuscript.

References

Bakheet T, Hitti E, Al-Saif M, Moghrabi WN, Khabar KSA. 2018. The AU-rich element landscape across human transcriptome reveals a large proportion in introns and regulation by ELAVL1/HuR. *Biochim Biophys Acta* **1861:** 167–177. doi:10.1016/j.bbagrm.2017.12.006

Cai L, Fritz D, Stefanovic L, Stefanovic B. 2010. Binding of LARP6 to the conserved 5′ stem-loop regulates translation of mRNAs encoding type I collagen. *J Mol Biol* **395:** 309–326. doi:10.1016/j.jmb.2009.11.020

Chen Y, Zubovic L, Yang F, Godin K, Pavelitz T, Castellanos J, Macchi P, Varani G. 2016. Rbfox proteins regulate microRNA biogenesis by sequence-specific binding to their precursors and target downstream Dicer. *Nucleic Acids Res* **44:** 4381–4395. doi:10.1093/nar/gkw177

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. 2011. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39:** D301–D308. doi:10.1093/nar/gkq1069

Damianov A, Kann M, Lane WS, Bindereif A. 2006. Human RBM28 protein is a specific nucleolar component of the spliceosomal snRNPs. *Biol Chem* **387:** 1455–1460. doi:10.1515/BC.2006.182

Darnell RB. 2010. HITS-CLIP: panoramic views of protein–RNA regulation in living cells. *Wiley Interdiscip Rev RNA* **1:** 266–286. doi:10.1002/wrna.31

Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van Nostrand EL, Pratt GA, et al. 2018. Sequence, structure, and context preferences of human RNA binding proteins. *Mol Cell* **70:** 854–867.e9. doi:10.1016/j.molcel.2018.05.001

Draper DE. 1999. Themes in RNA–protein recognition. *J Mol Biol* **293:** 255–270. doi:10.1006/jmbi.1999.2991

Farazi TA, Leonhardt CS, Mukherjee N, Mihailovic A, Li S, Max KE, Meyer C, Yamaji M, Cekan P, Jacobs NC, et al. 2014. Identification of the RNA recognition element of the RBPMS family of RNA-binding proteins and their transcriptome-wide mRNA targets. *RNA (New York, NY)* **20:** 1090–1102. doi:10.1261/rna.045005.114

Fu M, Blackshear PJ. 2017. RNA-binding proteins in immune regulation: a focus on CCCH zinc finger proteins. *Nat Rev Immunol* **17:** 130–143. doi:10.1038/nri.2016.129

Gerstberger S, Hafner M, Tuschl T. 2014. A census of human RNA-binding proteins. *Nat Rev Genet* **15:** 829–845. doi:10.1038/nrg3813

Glisovic T, Bachorik JL, Yong J, Dreyfuss G. 2008. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582:** 1977–1986. doi:10.1016/j.febslet.2008.03.004

Habacher C, Ciosk R. 2017. ZC3H12A/MCPIP1/Regnase-1-related endonucleases: an evolutionary perspective on molecular mechanisms and biological functions. *Bioessays* **39:** 1700051. doi:10.1002/bies.201700051

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, et al. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141:** 129–141. doi:10.1016/j.cell.2010.03.009

Haque N, Ouda R, Chen C, Ozato K, Hogg JR. 2018. ZFR coordinates cross-talk between RNA decay and transcription in innate immunity. *Nat Commun* **9:** 1145. doi:10.1038/s41467-018-03326-5

Hayakawa-Yano Y, Suyama S, Nogami M, Yugami M, Koya I, Furukawa T, Zhou L, Abe M, Sakimura K, Takebayashi H, et al. 2017. An RNA-binding protein, Qki5, regulates embryonic neural stem cells through pre-mRNA processing in cell adhesion signaling. *Genes Dev* **31:** 1910–1925. doi:10.1101/gad.300822.117

Hentze MW, Castello A, Schwarzl T, Preiss T. 2018. A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol* **19:** 327–341. doi:10.1038/nrm.2017.130

Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, Zhao BS, Mesquita A, Liu C, Yuan CL, et al. 2018. Recognition of RNA N^6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nat Cell Biol* **20:** 285–295. doi:10.1038/s41556-018-0045-z

Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE. 2004. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol* **11:** 257–264. doi:10.1038/nsmb738

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152:** 327–339. doi:10.1016/j.cell.2012.12.009

Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J. 2015. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527:** 384–388. doi:10.1038/nature15518

Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. 2001. Protein–RNA interactions: a structural analysis. *Nucleic Acids Res* **29:** 943–954. doi:10.1093/nar/29.4.943

Keck JL, Goedken ER, Marqusee S. 1998. Activation/attenuation model for RNase H: a one-metal mechanism with second-metal inhibition. *J Biol Chem* **273:** 34128–34133. doi:10.1074/jbc.273.51.34128

Korhonen JH, Palin K, Taipale J, Ukkonen E. 2017. Fast motif matching revisited: high-order PWMs, SNPs and indels. *Bioinformatics* **33:** 514–521.

Lai WS, Kennington EA, Blackshear PJ. 2002. Interactions of CCCH zinc finger proteins with mRNA: non-binding tristetraprolin mutants exert an inhibitory effect on degradation of AU-rich element-containing mRNAs. *J Biol Chem* **277:** 9606–9613. doi:10.1074/jbc.M110395200

Lambert NJ, Robertson AD, Burge CB. 2015. RNA Bind-n-Seq: measuring the binding affinity landscape of RNA-binding proteins. *Meth Enzymol* **558:** 465–493. doi:10.1016/bs.mie.2015.02.007

Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, Hu Z, Venkatesan K, Bethel G, Martin P, et al. 2007. hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89:** 307–315. doi:10.1016/j.ygeno.2006.11.012

Leppek K, Schott J, Reitter S, Poetz F, Hammond MC, Stoecklin G. 2013. Roquin promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs. *Cell* **153:** 869–881. doi:10.1016/j.cell.2013.04.016

Lin RJ, Chien HL, Lin SY, Chang BL, Yu HP, Tang WC, Lin YL. 2013. MCPIP1 ribonuclease exhibits broad-spectrum antiviral effects through viral RNA binding and degradation. *Nucleic Acids Res* **41:** 3314–3326. doi:10.1093/nar/gkt019

Liu Y, Beyer A, Aebersold R. 2016. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165:** 535–550. doi:10.1016/j.cell.2016.03.014

Loughlin FE, Mansfield RE, Vaz PM, McGrath AP, Setiyaputra S, Gamsjaeger R, Chen ES, Morris BJ, Guss JM, Mackay JP. 2009. The zinc fingers of the SR-like protein ZRANB2 are single-stranded RNA-binding domains that recognize 5′ splice site-like sequences. *Proc Natl Acad Sci* **106:** 5581–5586. doi:10.1073/pnas.0802466106

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci* **102:** 10557–10562. doi:10.1073/pnas.0409137102

Mackereth CD, Sattler M. 2012. Dynamics in multi-domain protein recognition of RNA. *Curr Opin Struct Biol* **22:** 287–296. doi:10.1016/j.sbi.2012.03.013

Martino L, Pennell S, Kelly G, Busi B, Brown P, Atkinson RA, Salisbury NJ, Ooi ZH, See KW, Smerdon SJ, et al. 2015. Synergic interplay of the La motif, RRM1 and the interdomain linker of LARP6 in the recognition of collagen mRNA expands the RNA binding repertoire of the La module. *Nucleic Acids Res* **43:** 645–660. doi:10.1093/nar/gku1287

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EE, et al. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4:** e04837. doi:10.7554/eLife.04837

Orenstein Y, Ohler U, Berger B. 2018. Finding RNA structure in the unstructured RBPome. *BMC Genomics* **19:** 154. doi:10.1186/s12864-018-4540-1

Pape UJ, Rahmann S, Vingron M. 2008. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics* **24:** 350–357. doi:10.1093/bioinformatics/btm610

Queiroz RML, Smith T, Villanueva E, Marti-Solano M, Monti M, Pizzinga M, Mirea DM, Ramakrishna M, Harvey RF, Dezi V, et al. 2019. Comprehensive identification of RNA–protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat Biotechnol* **37:** 169–178. doi:10.1038/s41587-018-0001-2

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Ray D, Kazan H, Chan ET, Peña Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR. 2009. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* **27:** 667–670. doi:10.1038/nbt.1550

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding

motifs for decoding gene regulation. *Nature* **499:** 172–177. doi:10.1038/nature12311

Rosario R, Childs AJ, Anderson RA. 2017. RNA-binding proteins in human oogenesis: balancing differentiation and self-renewal in the female fetal germline. *Stem Cell Res* **21:** 193–201. doi:10.1016/j.scr.2017.04.008

Senissar M, Manav MC, Brodersen DE. 2017. Structural conservation of the PIN domain active site across all domains of life. *Protein Sci* **26:** 1474–1492. doi:10.1002/pro.3193

Skrisovska L, Bourgeois CF, Stefl R, Grellscheid SN, Kister L, Wenter P, Elliott DJ, Stevenin J, Allain FH. 2007. The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. *EMBO Rep* **8:** 372–379. doi:10.1038/sj.embor.7400910

Stormo GD. 1988. Computer methods for analyzing sequence recognition of nucleic acids. *Annu Rev Biophys Biophys Chem* **17:** 241–263. doi:10.1146/annurev.bb.17.060188.001325

Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, NY)* **249:** 505–510. doi:10.1126/science.2200121

Varani G, McClain WH. 2000. The G x U wobble base pair: a fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep* **1:** 18–23. doi:10.1093/embo-reports/kvd001

Weick EM, Puno MR, Januszyk K, Zinder JC, DiMattia MA, Lima CD. 2018. Helicase-dependent RNA decay illuminated by a Cryo-EM structure of a human nuclear RNA exosome–MTR4 complex. *Cell* **173:** 1663–1677.e21. doi:10.1016/j.cell.2018.05.041

Weyn-Vanhentenryck SM, Mele A, Yan Q, Sun S, Farny N, Zhang Z, Xue C, Herre M, Silver PA, Zhang MQ, et al. 2014. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep* **6:** 1139–1152. doi:10.1016/j.celrep.2014.02.005

Wilamowski M, Gorecki A, Dziedzicka-Wasylewska M, Jura J. 2018. Substrate specificity of human MCPIP1 endoribonuclease. *Sci Rep* **8:** 7381. doi:10.1038/s41598-018-25765-2

Xia H, Chen D, Wu Q, Wu G, Zhou Y, Zhang Y, Zhang L. 2017. CELF1 preferentially binds to exon-intron boundary and regulates alternative splicing in HeLa cells. *Biochim Biophys Acta* **1860:** 911–921. doi:10.1016/j.bbagrm.2017.07.004

Xu J, Peng W, Sun Y, Wang X, Xu Y, Li X, Gao G, Rao Z. 2012. Structural study of MCPIP1 N-terminal conserved domain reveals a PIN-like RNase. *Nucleic Acids Res* **40:** 6957–6965. doi:10.1093/nar/gks359

Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, et al. 2017. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (New York, NY)* **356:** eaaj2239. doi:10.1126/science.aaj2239

Received November 7, 2019; accepted in revised form June 23, 2020.

Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences

Arttu Jolma, Jilin Zhang, Estefania Mondragón, et al.

Genome Res. 2020 30: 962-973 originally published online July 23, 2020
Access the most recent version at doi:10.1101/gr.258848.119